CIC 灼识

全球大模型行业报告

摘要

全球大模型产业已告别实验探索期，正式跃升为驱动全球智能化转型的核心引擎。大模型凭借卓越的扩展性与泛化能力，正为复杂的认知任务打造一套标准化的底层智能架构。

目录

1. 行业概览

1.1 行业定义

全球大模型行业是人工智能领域极具变革性的细分赛道，更是推动全球社会智能化变革的核心引擎。这类模型释放出前所未有的生产力与认知创造力，持续重新定义人类潜能的边界。与受限于特定化场景的传统小模型不同，大模型从设计之初便具备与生俱来的扩展性与卓越的泛化能力。

大模型技术公司作为行业的核心创新主体，可进一步划分为以下两类：

Pureplay（即主营业务聚焦于大模型）：企业核心资源、技术积淀与商业模式均完全围绕大模型的研发及商业化布局，通过资源的集中投入推动技术快速创新。

Non-pureplay（即在原有业务基础上切入大模型的公司，如大型互联网平台、云计算服务商等）：大型互联网平台及云计算服务商依托资本与算力优势，将大模型技术融入自身产品生态，加速技术验证与商业化落地。

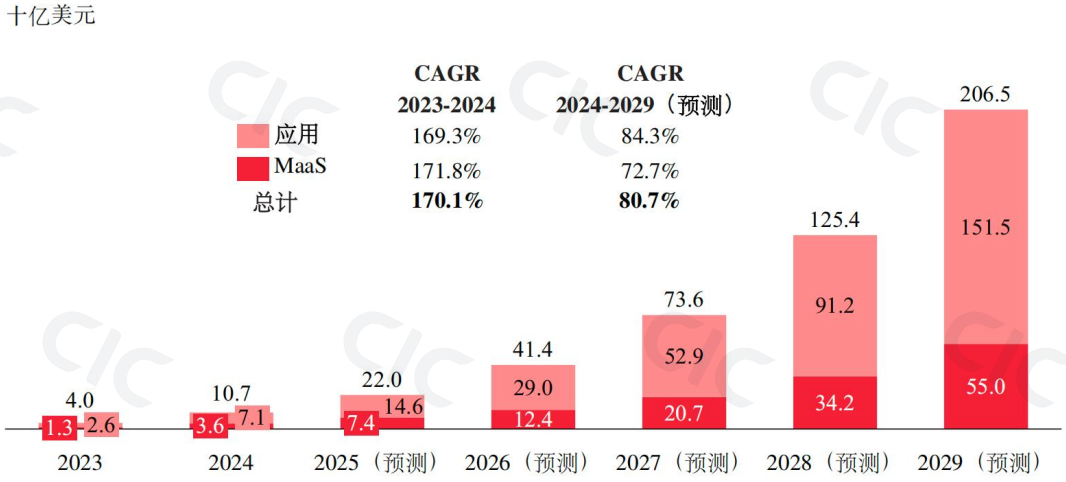
1.2 行业规模与增长

全球大模型市场的收入来源分为两类，即基于模型的收入与基于部署的收入，其中前者是市场增长的核心引擎。基于模型的收入主要来自两方面，一是通过订阅模式产生的 AI 原生应用收入，二是通过云端 API 及授权方式实现的 MaaS（model-as-a-service）收入。而基于部署的收入则来源于定制化的私有化部署解决方案。

在技术日趋成熟、用户付费意愿持续提升的双重驱动下，全球大模型市场（按基

于模型的收入口径计）迎来爆发式增长。据 CIC 灼识咨询测算，2024 年该市场规模为 107 亿美元，预计 2029 年将增至 2,065 亿美元，年复合增长率（CAGR）达 80.7%。其中，大模型应用市场 2029 年规模预计将达 1,515 亿美元，大模型 MaaS 市场同期规模将增至 550 亿美元。

全球大模型市场规模，按基于模型的收入计，**2023-2029**（预测）



资料来源：CIC 灼识

附注：基于模型的收入主要包括来自大模型应用订阅服务以及大模型 API 调用和授权的收入。

2. 核心增长驱动因素与发展趋势

2.1 核心驱动因素

技术跃升

全球大模型行业的发展以颠覆性技术突破为核心特征，每一代模型的迭代都解锁了前所未有的应用边界与商业价值。值得关注的是，“交错思维”框架、Claude 3.7 的代码能力升级以及 Claude Code 这类创新成果，推动行业范式实现根本性转变，从被动响应的工具形态，升级为可自主完成任务协同的主动式 AI 智能体。

Scaling Law

Scaling Law 仍是支撑行业增长的核心底层驱动因素。预训练阶段的 Scaling Law——即模型性能随模型规模、数据量与算力投入的提升而提升——在文本、音频、视频领域均依然适用。

当今，全新的推理时算力 Scaling Law 已然出现：2025 年头部推理模型的表现证明，推理算力投入越大，模型智能水平越高。预训练与推理阶段的规模拓展形成协同效应，正在为行业构筑一条全新的“摩尔定律”，推动模型吞吐量与复杂问题解决能力的整体进步。

成本下降与市场落地

相较于模型能力提升，推理成本的持续下降是更为确定的行业驱动因素。推理成本已从 2022 年末约每百万 token 20 美元，大幅降至 2024 年末的 0.1 美元以下。

在架构创新、工程优化及算力成本下降的共同驱动下，推理成本预计每年以十倍幅度下降，让此前不具备可行性的垂直领域应用（如大规模内容审核、实时 AI 陪伴）实现商业化落地。这一成本优化趋势大幅降低了应用门槛，加速推动大模型在海量工业及消费场景中实现广泛渗透与融合。

2.2 核心发展趋势与竞争壁垒

模型智能水平持续提升

模型规模扩展与能力提升

基础模型的参数量实现大幅跃升，性能亦取得显著提升，GPT 系列模型在专业评测中展现出接近人类水平的推理与理解能力。混合专家（MoE）架构成为核心技术突破，在扩大模型规模的同时，实现了对算力成本与推理延迟的有效管控。

上下文长度与推理效率提升

现代基础模型已从 GPT-3 的 2,048 token 限制，升级至数百万 token 的上下文窗口，可实现与超长文档的高保真交互。但更长的上下文窗口推高了推理成本，进而催生了架构创新与检索增强生成技术。其中最具代表性的突破，便是对注意力机制的优化。

人类对齐

RLHF（基于人类反馈的强化学习）已成为基础模型的标准训练流程，有效提升了模型对用户指令的遵从度与回复质量。人类对齐模型在准确率、语气控制以及处理不当查询等方面均实现了显著优化。

CoT（思维链）与推理模型涌现

2022 年推出的 CoT（思维链）提示技术，有效提升了模型在复杂推理任务上的表现。2024 年行业迎来重大转变：模型在推理阶段被训练为逐步拆解问题，并将更多算力投入到迭代推理、反思与输出优化中。推理是一项可计算的过程，而非仅由模型规模自然涌现。此外，成本、延迟与效果的权衡将使未来模型形成分化：一类专注于快速、精准的响应；另一类则用于深度、高算力消耗的推理，而推理时算力将成为核心关键。

Agentic 工具使用成为新范式

AI Agent 已成为全新的发展范式，使模型能够自主规划并借助外部工具完成复杂任务。2023 年，GPT-4 的插件与函数调用功能打破了模型仅依托训练数据运行的局限，Gemini 则实现了代码在沙箱环境中的自主运行。2025 年，头部企业进一步强化了模型的 Agent 能力，使其从被动应答者转变为主动的任务统筹者。

闭源与开源共进

近年来，闭源与开源模型实现了并行发展。开源趋势推动闭源研发方加快迭代速度，同时也为用户提供了更丰富的定制化选择。

加速发展

全球基础模型的智能水平持续提升。根据 OpenAI 的五级路线图，当前模型已达到 Level 3 交界处。展望未来，行业发展轨迹清晰指向加速进。

级别	名称	描述
L1	Chatbots	可对话的AI
L2	Reasoners	具备像人类一样解决问题能力的AI
L3	Agents	能行动的AI
L4	Innovators	能辅助发明的AI
L5	Organizations	能像组织一样运作的AI

资料来源: OpenAI

模态持续扩充

从单模态到多模态

基础模型已拓展至多模态领域，旨在将文本、图像、音频、视频等特征整合并对齐至统一语义空间，进而实现跨模态融合。

视觉理解

多模态理解仍处于发展初期，但近期行业趋势正转向打造更统一的多模态能力。以 GPT-4V 为例，它在 GPT-4 框架基础上完成扩展，支持图像输入，可让用户指令模型分析视觉内容、描述图像细节、解读网络梗图的趣味内涵以及医学影像中的信息。Gemini 基于纯解码器架构构建，支持图像、视频、音频多种模态，并在多模态推理任务中创下全新性能基准。

语言生成

文本-音频融合让 AI 具备了音频理解与生成能力。2023 年音频合成技术实现飞

速突破：ElevenLabs 与 MiniMax 推出了高度拟人化的语音效果，OpenAI 则为 ChatGPT 加入实时语音对话能力，大幅推动了其在智能助手、客服场景的应用。这一融合技术也催生了语音 AI 智能体与智能硬件设备。未来模型将能更精准地捕捉语音中的情绪与意图，给出更自然的应答，实现更流畅的人机交互。

视觉生成

2022 年，扩散模型与 Transformer 架构的融合，实现了从文本提示生成高保真、艺术化图像的能力。2023 年以来，这一技术趋势转向视频生成，简化了各行业的内容生产流程，大幅提升创作效率。当前前沿研究正探索统一多模态模型，将大语言模型的自回归推理能力与扩散模型的精准生成特性相结合，以此复刻人类智能的融合式认知特点。

大模型原生产品渗透率持续提升，商业价值不断释放

大模型应用增速迎来前所未有的发展

过去三年，下一代人工智能迎来爆发式增长，增速超越历史上所有技术浪潮，

相关 AI 产品在用户规模与商业收入上均实现极速扩张。在现有技术基础设施的支撑下，人工智能正快速渗透至互联网全域，人类正处于技术指数级增长的关键转折点。尽管从实时视角看，技术发展看似线性推进，但指数级的飞跃往往会在短时间内集中到来。

通用性为大模型应用带来广阔的市场空间

基础模型正通过单一可扩展架构实现大规模部署与个性化应用，彻底变革生产力、娱乐及企业级服务领域。这种通用性让模型能够覆盖从全球企业到个人创作者的广泛用户群体，实现超高的投入产出比。

2.3 未来展望

基础模型应用的商业化仍处于初期阶段，而智能体集成是释放其巨大商业价值的关键转折点。

Agent 应用: AI 智能体正从工具提供演变为端到端价值交付者，正在将市场重心从企业软件领域，转向规模达数万亿美元的全球劳务服务市场。

娱乐及生成类应用: 在沉浸式共创需求的驱动下，搭载情感智能的个性化 AI 陪伴伙伴，正重新定义娱乐行业格局，建立更深度的用户联结。AI 视频生成让内容创作走向平民化，高级语音接口成为多模态交互的标准形态，进一步推动了这一转变。而这些趋势共同简化了人机沟通，助力各垂直领域实现加速增长。

多模态应用: 2025 年 3 月，GPT-4o 的原生多模态升级显著提升了图像质量并推动订阅用户增长，充分验证了集成式多模态技术的商业潜力。该模型实现了精准的图像内文本生成与细粒度编辑功能，为教育、营销和科学插画等领域解锁了专业级应用场景。未来，文本、音频与视觉的深度融合将催生完全可编辑、同步化的视频创作，进而驱动全球短视频领域的下一次革命性变革。

值得注意的是，在 Scaling Law 与推理成本大幅下降的双重驱动下，基础模型已不再是遥不可及的前沿技术。随着人工智能向金融、制造等垂直行业深度渗透，

行业竞争将愈发聚焦于三大核心壁垒：前沿研发能力、商业化效率、以人才为核心的组织实力，这也将决定下一代行业领军者的格局。

未来，全球基础模型产业将汇聚成成熟的智能体驱动型生态，成为人类社会不可或缺的核心基础设施，从根本上重塑全球生产体系，进一步夯实其作为新一轮生产力革命核心引擎的地位。

关于 CIC 灼识

CIC 灼识咨询（简称“CIC 灼识”）是一家专业咨询机构，围绕投融资全生命周期，提供定制化一站式全流程服务。公司在全球各大市场主导打造多个行业首创的标杆 IPO 项目，业绩稳居世界前列。同时在各类专业细分赛道中，拥有无可匹敌的资源触达能力与深度全覆盖研究实力。

CIC 灼识助力企业优化具备规模化潜力的商业模式，塑造极具说服力的资本市场价值叙事，畅通对接全球资本市场的路径。同时作为投资机构信赖的尽职调查合作伙伴，输出精细化行业研判视角，并直通各领域权威专家资源，助力客户精准锁定高价值机遇、有效规避核心重大风险。

CIC 灼识团队深耕金融服务、人工智能、大数据、互联网、高新技术、医疗健康、教育、文娱、消费品、交通运输物流、能源电力、环境与建筑科技、化工、工业制造、农业等多元领域，实时掌握深度一线市场动态，能够为客户独家输出贴合细分行业、可落地执行的专业洞察结论。

CIC 灼识报告 & 行业概览

CIC 灼识搭建了一套严谨的多元化研究框架，整合一手调研与二手资料，为所有分析研判筑牢根基。一手调研主要深度对接行业权威专家与一线从业从业者，重点深耕供应链金融领域。二手研究则汇总梳理各大权威机构的公开数据，数据来源包括：中华人民共和国国家统计局、国家金融监督管理总局（SAFR，原中国

银行业监督管理委员会)、中国证券监督管理委员会，以及上市公司公开披露文件。

我们运用自研专属数据分析体系对收集到的信息进行加工处理，并通过多渠道研究数据交叉比对验证研究结论，确保分析过程严谨、结果真实可靠。

本报告中展示的所有统计数据均可核验追溯，全部基于报告出具当日可获取的有效信息整理而成。

本篇内容摘编提炼自 CIC 灼识深度行业研究报告精华，聚焦各细分赛道的供需走势、核心增长驱动因素、研发创新趋势与行业未来发展前景等核心内容，同时融合专家访谈、市场实地调研、行业数据解析等多维度专业研判成果。

免责声明

本报告由 CIC 灼识依据截至出具当日可获取的信息编制。本报告仅作信息参考之用，内容不具备最终定论效力，亦不得被解读为确定性结论。

本报告所载全部内容，均不构成且不得视作投资建议、投资推荐，亦非开展任何投资活动的要约、招揽或劝导。

凡因使用或依赖本报告所载信息，直接或间接引发任何损失、损害及各类索赔诉求的，CIC 灼识特此明确免除一切相关责任。

联系我们

如需了解本报告更多详情，或咨询 CIC 灼识的各项专业服务，欢迎访问 CIC 灼识官方网站，亦可发送邮件至：marketing@cninsights.com。